Digitization Pilot Project - Final Report
January 2003

Project Goal

The overarching goal of the pilot has been to develop an information structure and process for identifying, evaluating, digitizing, and enhancing discovery of foreign language consumer health information resources (heretofore only available as hardcopy documents) on the web.

Identifying Resources

At the start of the project, a librarian consultant hired by PSRML used data from the Region 7 Consumer Health Library Directory survey to identify potential library sources of foreign language consumer health information resources in the NN/LM Pacific Southwest Region. She contacted these sources, obtained documents, and sent the documents along with a report of her overall findings to PSRML. Even though some libraries had reported foreign language collections, the number and scope of documents ultimately identified was quite limited. Because of this, PSRML decided to scale down the project and to do it "in-house" rather than to hire a project manager, on a consultant basis, to conduct the project.

Evaluation of Resources

Information about quality was not easy to determine from the set of documents identified for the pilot project. It was not always possible to find out the original source of content for a document; the age of the content, the date of publication, or about rights to reproduce the content. Many agencies do not have explicit review standards for the materials they produce, for quality of translation/cultural appropriateness, or for readability of these materials. While we could apply information quality filters in reviewing documents; we were not in a role to weigh in on content issues (see Evaluation Criteria).

Digitization Process

Prior to digitizing documents, PSRML staff attended a workshop on database-driven websites, and met with Howard Batchelor, Digital Library Coordinator at UCLA, about the feasibility of using tools already developed at UCLA to provide access to digitized materials. Because digitization efforts at UCLA have been driven by preservation and archival concerns, PSRML staff had some difficulty in interpreting and adapting the tools to the more immediate, and practical needs of the pilot. A sample set of documents from a variety of organizations was created; individual documents were scanned and saved as

.pdf files. Descriptive item level data and digital object data were captured and entered into the UCLA Digital Library Core Database (CD)--an MSAccess database. Equipment used to digitize the documents included an HP ScanJet 4c and an IBM compatible PC. Software included Adobe Acrobat 5.0, Adobe Photoshop 6.0, and DeskScan II 2.9 (scanning software). After data was entered into the CD, the data and the database file structure were exported into a MySQL database. Each document was scanned in two ways to facilitate viewing and printing. The image created for viewing was in the logical sequence of the pages to be read by the user. The image created for printing was in a sequence that, if printed using a duplex printer, recreated a version of the original document. The resolution of 100 pixels per inch was decided by taking into account both image file size and printing quality. Images were stored on the UCLA Biomedical Library/PSRML server. Ultimately, a mock-up of the project website design was produced and housed on the NN/LM server. The original intent of the project was to concentrate on consumer health information materials in languages other than Spanish or English. However, to explore metadata issues, it was easiest to work with a sample set of documents that were in English with Spanish translations.

Metadata

A metadata scheme was produced, incorporating collection level data, digital object data, and administrative data from the CD. The scheme also includes discovery data, incorporating Dublin Core elements, qualifiers, and encoding schemes found in the CD and in other project tools accessible via the web; namely, tools linked from the Dublin Core Metadata Initiative site (www.dublincore.org). Subject headings and keywords were applied to individual documents.

Evaluation

Because the project was scaled down, not all of the original project objectives were addressed. The extent, characteristics, and location of foreign language consumer health information resources in Los Angeles and San Francisco were not explored beyond what was identified by the consultant. In retrospect, it might have been better to target potential content producers rather than libraries in locating resources (i.e. public health departments, health systems, commercial providers, etc.). Quality standards were drafted for the project; they may serve as a useful starting point for further discussion before embarking on future projects of this nature. The process of digitization was successfully developed and tested, although on a smaller scale and over a much longer period of time than originally proposed, because of the limited amount of staff time that could be devoted to the project.

Recommendations

Recommendations to NLM, regarding its role in assessing need and in providing access to multilingual/multicultural consumer health information resources, were forwarded via

the Multilingual/Multicultural Materials Working Group of the NN/LM. Hopefully, this project has laid the groundwork for future projects to enhance the public's access to these resources.

For More Information

If you have any questions, please contact Heidi Sandstrom or Andrea Lynch at:

Pacific Southwest Regional Medical Library
UCLA Louise M. Darling Biomedical Library
12-077 Center for the Health Sciences
Box 951798
Los Angeles, CA 90095-1798
(310)825-1200 or (800)338-7657; psr-nnlm@library.ucla.edu

Digitization Pilot Project
January 2003

Administrative Metadata

Collection Level Data

ParentDivID - PSRML
Object Type – Collection
Label – Consumer Health Materials (non book, non serial)
DivID – 00001

Item Data

Scanning Data

File Name
Image size
Document size
Resolution (low)
Resolution (high)
OCR (?) – may want to make Word doc available
Scanning notes
Printing – paper size
Assembly instructions

Cataloging Data

Identifier on document (e.g. Title #R802; CAC M101)

Digitization Pilot Project
January 2003

Metadata Scheme

References:

Dublin Core Metadata Element Set, Version 1.1: Reference Description (see http://www.dublincore.org/documents/dces/)
Dublin Core Qualifiers
(see http://dublincore.org/documents/dcmes-qualifiers/)

Element: Title

DC Qualifier: Alternative

Name: alternative
Label: Alternative
Definition: Any form of the title used as a substitute or alternative to the formal title of the resource.
Comment: This qualifier can include Title abbreviations as well as **translations**.

Element: Creator

Element: Subject

Encoding Schemes for *Subject*:

Name: MeSH
Label: MeSH
Definition: Medical Subject Headings
See also: http://www.nlm.nih.gov/mesh/meshhome.html

[Subject Headings used by MEDLINEplus]
[Subheading/aspect (i.e. prevention, diagnosis, treatment)]

Element: Description

DC Qualifier: Abstract

Name: abstract
Label: Abstract
Definition: A summary of the content of the resource.

Element: Publisher

Element: Contributor (to content – e.g. illustrator, other than creator, translator)

Element: Date (YYYY-MM-DD)

DC Qualifier: Created

Name: created
Label: Created
Definition: Date of creation of the resource.

Translated
DC Qualifier: Valid (date range)

Name: valid
Label: Valid
Definition: Date (often a range) of validity of a resource.

DC Qualifier: Issued (publication)

Name: issued
Label: Issued
Definition: Date of formal issuance (e.g., publication) of the resource.

Placed Online*
Record Created (metadata record)*
Translation

Encoding Schemes for *Date*:

DCMI Period

Name: Period
Label: DCMI Period
Definition: A specification of the limits of a time interval.
See also: http://dublincore.org/documents/dcmi-period/

W3C-DTF

Name: W3CDTF
Label: W3C-DTF
Definition: W3C Encoding rules for dates and times - a profile based on ISO 8601
See also: http://www.w3.org/TR/NOTE-datetime

Element: Type

Encoding Schemes for *Resource Type*:

DCMI Type Vocabulary

Name: DCMIType
Label: DCMI Type Vocabulary
Definition: A list of types used to categorize the nature or genre of the content of the resource.
See also: http://dublincore.org/documents/dcmi-type-vocabulary/

Text (includes images of text)/PDF

Element: Format

Qualifiers that refine *Format*:

Extent

Name: extent
Label: Extent
Definition: The size or duration of the resource.

Medium

Name: medium
Label: Medium
Definition: The material or physical carrier of the resource. (e.g. single sheet pamphlet, bookmark, postcard, booklet, tape, video)

Encoding Schemes for *Format*:

IMT

Name: IMT
Label: IMT

Definition: The Internet media type of the resource.
See also: http://www.isi.edu/in-notes/iana/assignments/media-types/media-types

[Platform (what's needed to access; e.g. Adobe Acrobat)]

Element: Identifier (unique number for physical item)

Encoding Schemes for *Resource Identifier*:

URI

Name: URI
Label: URI
Definition: A URI Uniform Resource Identifier
See also: http://www.ietf.org/rfc/rfc2396.txt

Element: Source

Element: Language

Encoding Scheme for *Language*:

RFC 1766

Name: RFC1766
Label: RFC 1766
Definition: Internet RFC 1766 'Tags for the identification of Language' specifies a two letter code taken from ISO 639-2, followed optionally by a two letter country code taken from ISO 3166.
See also: http://www.ietf.org/rfc/rfc1766.txt

Element: Relation

isTranslationOf
hasTranslation
isOrderInfo
isAgencyReview
isUserReview

isReplacedBy

Name: isReplacedBy
Label: Is Replaced By

Definition: The described resource is supplanted, displaced, or superceded by the referenced resource.

Replaces

Name: replaces
Label: Replaces
Definition: The described resource supplants, displaces, or supersedes the referenced resource.

isFormatOf

Name: isFormatOf
Label: Is Format Of
Definition: The described resource is the same intellectual content of the referenced resource, but presented in another format.

hasFormat

Name: hasFormat
Label: Has Format
Definition: The described resource pre-existed the referenced resource, which is essentially the same intellectual content presented in another format.

isSponsoredBy

Element: Coverage

Spatial (i.e. national, state, county, city, etc.)

Element: Rights

Rights management statement
Price Code

*Non-Dublin Core Elements

Element: Quality*

Element: Audience*

Mediator (e.g. tool for nurse, health educator, etc.)
Beneficiary

Level (e.g. literacy level 3rd grade)
Age
Prerequisites

Element: Cataloger*

Digitization Pilot Project
January 2003

Website design

Homepage – Top Level

Title
Brief annotation
Search (by language, subject, keyword)
About contributors of collections
About NN/LM
About project/website

Search

Language	A-L dropdown M-Z dropdown
OR	
Subject	A-L dropdown M-Z dropdown
OR	
Keyword	Box with instruction for phrases, multiple terms Boolean operators

Filters

Aspect	dropdown (Prevention, Diagnosis, Treatment)
Audience	checkboxes

Language Retrieval Display – 2nd Level

Example: Language (e.g. chuukese)
Subject (alpha)
Title (alpha) – links to record display

Subject Retrieval Display – 2nd Level

Example: Subject (e.g. communicable diseases)
Language (alpha)
Title (alpha) – links to record display

Keyword Retrieval Display – 2nd Level

Example: Title 1 – links to record display
Title 2 – links to record display
etc.

Record Retrieval Display – 3rd Level

Title – links to online view of item OR to print item OR to vendor
Reading Level
Audience
Subjects
Keywords
Resource Type (e.g. pamphlet, booklet, video)
Format: text/pdf
Cost
Cataloging Agency
See What Users Say (link)
Comments/Recommendations for Improving This Item

Contribute pamphlet – 2nd Level

Form (see metadata schema)
e-submission of data
Verification/Confirmation
Instructions (print confirmation and mail with pamphlet, video, etc.)

Digitization Pilot Project

Evaluation Criteria

- Authoritative/reliable source for information
- Content is up-to-date, accurate
- Purpose to inform, not to sell product or service (not for referral or promotion only)
- Content suggests an action/behavior
- Material supports health professional/patient relationships
- Material avoids bias/stereotyping concerning women and ethnic groups (both text and graphics)
- Rights information is implicit or explicitly stated (e.g. government publication; states "You may reproduce this…"; has © symbol and date)
- Content is readable; sentence structure/vocabulary appropriate for general public
- Is age appropriate in text/graphics (e.g. if for children, simple text and suitable graphics, plays on interests)

Design issues:

Has clear message
Logical design
Well-organized

Content Producer:

Do you know about the review standards of the organization/agency? Does it use a rating scale for materials?